Improving the Action Research Arm test: a unidimensional hierarchical scale

van der Lee, J.H.; Roorda, L.D.; Beckerman, H.; Lankhorst, G.J.; Bouter, L.M.

published in
Clinical Rehabilitation
2002

DOI (link to publisher)
10.1191/0269215502cr534oa

document version
Publisher's PDF, also known as Version of record

Improving the Action Research Arm test: a unidimensional hierarchical scale

Johanna H van der Lee, Leo D Roorda, Heleen Beckerman, Gustaaf J Lankhorst and Lex M Bouter

Clin Rehabil 2002 16: 646
DOI: 10.1191/0269215502cr534oa

Clinical Rehabilitation 2002; **16**: 646–653

Improving the Action Research Arm test: a unidimensional hierarchical scale

Johanna H van der Lee, Leo D Roorda, Heleen Beckerman, Gustaaf J Lankhorst Department of Rehabilitation Medicine and Institute for Research in Extramural Medicine and **Lex M Bouter** Institute for Research in Extramural Medicine, Vu University Medical Center, Amsterdam, The Netherlands

Received 20th September 2000; returned for revisions 4th January 2001; revised manuscript accepted 7th June 2001.

Background: The Action Research Arm (ARA) test is a performance test of upper extremity motor function which consists of 19 items divided into four hierarchical subtests. This multidimensionality has not yet been tested empirically.
Objective: To investigate the dimensionality of the ARA test.
Design: Cross-sectional study involving a sample of 63 chronic stroke patients.
Methods: A Mokken scale analysis was performed.
Results: The Mokken scale analysis revealed one strong unidimensional scale containing all 19 items, of which the scalability coefficient *H* was 0.79, while *H* per item ranged from 0.69 to 0.86. The reliability coefficient rho equalled 0.98, indicating a very high internal consistency. A subset of 15 out of 19 items showed an invariant hierarchical item-ordering.
Conclusion: The ARA test is a unidimensional scale. The use of subtests, as proposed in the original description of the instrument, is not supported by the present findings. The 15-item scale presented here can be used for adaptive testing, i.e. using only a selected subset of items based on prior knowledge about the patient's abilities, thus minimizing testing time.

Introduction

The Action Research Arm (ARA) test is a performance test for upper extremity function and dexterity.[1] It was constructed by Lyle, and derived from the Upper Extremity Function Test (UEFT).[2] The main reasons why Lyle decided to change the UEFT and to construct the ARA test, were: (1) the time needed to complete the UEFT, which consisted of 33 items, (2) the complexity of certain items, and (3) perceived redundancy of other items, notably those involving repetitive fine finger–thumb oppostion movements.[1] Since Lyle's publication of the ARA test in 1981, its validity and reliability have been reconfirmed,[3–6] and this test has been used as an outcome measure in a number of clinical studies.[4,7–15] The responsiveness of the ARA test, which is another important clinimetric characteristic, in addition to reliability and validity, has been shown to be adequate in the first eight weeks post stroke[3] as well as in chronic stroke patients undergoing forced use therapy.[15,16]

Address for correspondence: JH van der Lee, Department of Rehabilitation Medicine, Vu University Medical Center, PO Box 7057, 1007 MB Amsterdam, The Netherlands. e-mail: jh.vanderlee@vumc.nl

10.1191/0269215502cr534oa

For practical purposes, the time required to administer a test should be as limited as possible. If all 19 items of the ARA test are performed, administration of the test takes about 20 minutes. Lyle grouped the 19 tasks (items) of the ARA test into four subtests, each of which was intended to constitute a hierarchical Guttman scale. In a Guttman scale, all items are ordered according to ascending difficulty.[17,18] If a patient can perform a particular task, this predicts his or her ability to perform all easier tasks. On the other hand, failure to perform a certain task predicts failure on all tasks which are more difficult. A test which meets the requirements of a Guttman scale, can be used for adaptive testing. This means that the score of an individual can be assessed by applying only a few items of the test, because the ability to perform all easier items is predicted by a 'success', and the inability to perform all more difficult items is predicted by a 'failure'.

For each subtest, Lyle described decision rules (see Appendix).[1] The patient must first try to perform the most difficult task in a subtest. If the maximum score (3 points) is obtained for this task, the maximum score is assigned to this subtest, and the patient proceeds to the most difficult task of the next subtest. If the patient does not obtain the maximum score on the most difficult task, the easiest task of the same subtest is tried. If the patient fails on the easiest task (score 0), the score on this subtest is 0. Only if the patient does not obtain the maximum score on the most difficult task and succeeds (partially) in performing the easiest task, does he or she have to try all intermediate items in the subtest. In this way, the number of tasks to be performed can be reduced from 19 (maximum) to a minimum of 4 (if the patient performs the most difficult task of each subtest normally).

From Lyle's original article,[1] the methods used to construct the subtests and to define the hierarchy within the subtests are not entirely clear. Lyle's description of 'trial and error' used to constitute the subtests does not suggest a rigorous, easily replicable scientific methodology. To our knowledge, the use of the four subtests has never been formally questioned.

The ARA test was used as a primary outcome measure in a randomized clinical trial to evaluate the effect of forced use treatment in chronic stroke patients.[15] Every patient was asked to attempt every task, so the decision rules stated by Lyle were not applied. The data obtained from the first baseline measurement were analysed in order to answer the following principal research question:

1) Can the subtests and decision rules described by Lyle be confirmed empirically by the obtained data? This was operationalized in two ways:
   a) In what proportion of cases does application of Lyle's decision rules lead to a sum-score which is different from the sum-score obtained by applying all 19 items?
   b) Do the 19 items of the ARA test constitute one or multiple dimensions (subtests)?

If Lyle's decision rules are not empirically confirmed in the data set, we will also answer the following question:

2) Can the ARA test be shortened in a different, methodologically sound and replicable way?

Finally, if the answer to the second question is positive, we will use the data from the second baseline measurement and the post-treatment measurement in the experimental group to compare the responsiveness of the different versions:

3) Is the responsiveness ratio different for the Lyle version, the 19-item version and the new, shortened version?

Methods

The first baseline measurement in the randomized clinical trial was obtained from 63 patients.[15] To be eligible for the trial, patients had to comply with the following inclusion criteria: (1) a history of a single stroke, at least one year previously, resulting in a hemiparesis on the dominant side; (2) a minimum of 20 degrees of active extension in the wrist and 10 degrees of finger extension; (3) Action Research Arm test score below 51 (maximum score 57); (4) age 18–80 years; (5) able to walk indoors without a stick,

indicating no major balance problems; (6) no severe aphasia (score above P50 on the SAN test (Stichting Afasie Nederland)[19]); (7) no severe cognitive impairments (Mini-Mental State Examination[20] score of 22 or higher). The protocol was approved by the hospital's Medical Ethical Committee, and all patients gave written informed consent.

Analysis

Summation of the scores of all 19 items of the ARA test yields a sum-score which ranges from 0 (none of the movements can be performed) to 57 (all tasks are performed normally). The proportion of patients whose sum-scores would have been different (higher or lower) from the 19-item sum-score if Lyle's decision rules had been followed was visualized by plotting the differences between the 'Lyle sum-score' and the '19-item sum-score' on the *y*-axis against the '19-item sum-score' on the *x*-axis, resulting in a (slightly modified) Bland–Altman plot.[21]

Mokken scale analysis was used to analyse the data (MSPWIN 5.0).[22] Mokken scale analysis uses a probabilistic approach, as opposed to the deterministic approach of the Guttman scale analysis.[23] It can be viewed as a nonparametric approach to the item response theory. Contrary to the Guttman approach, which can only be used for dichotomous (pass/fail) items, Mokken scale analysis can be used for polytomous items, which have more than two possible scores per item.[22]

The first model tested in Mokken scale analysis is the monotone homogeneity model, which means that: (1) items form a unidimensional scale (measuring the same construct or latent trait), (2) item scores are locally independent (meaning that item scores are independent within a group of persons with the same value of the latent trait), and (3) the item response function for each item is a monotonely nondecreasing function of the latent trait. If, in addition, (4) the item response functions do not intersect, the item-set is consistent with the double monotonicity model, the second model tested, indicating that the items have an invariant hierarchical ordering across the latent trait scale.[22]

The scalability coefficient H is a measure of the accuracy of ordering persons by means of the sum-score. The minimum value of scale H indicating a 'strong scale' is 0.50.[23,24] In a Guttman scale, H equals 1. Individual items do not fit in the scale if item $H < 0$. The internal consistency of a scale consistent with the doubly monotone model is indicated by rho, which can be interpreted as the item response theory-based equivalent of Cronbach's alpha.[22]

For a more detailed check of the assumptions of both models, the so-called Crit values are used. A scale is considered to adequately meet these assumptions if the largest Crit value per item for each assumption is smaller than 40. If the least favourable Crit value exceeds 80, this casts serious doubt on the validity of the model for this item.[22] Thus, Crit values can be used to delete 'less fitting' items from the scale. To answer the research question of uni- or multidimensionality (question 1b), all 19 items were included in the analysis. Subsequently, an item-set was constructed which was consistent with the double monotonicity model (question 2) by stepwise omission of the item with the highest Crit value of the nonintersection criterion, until a scale was obtained in which the highest Crit value was well below 80.

To compare the responsiveness of the Lyle version, the 19-item version and the new, shortened version (question 3), we used the data obtained from the experimental group, who underwent an intensive forced use treatment during two weeks, five days a week, 6 hours a day.[15] The responsiveness ratio was computed as the ratio of the mean change after the experimental intervention and the standard deviation of the mean change during the two-week baseline period.[16,25] Because all included patients were in the chronic phase, their arm function was considered to be stable during the baseline period.

Results

The baseline characteristics of the 63 patients are presented in Table 1. The ARA sum-scores based on all 19 items ranged from 5 to 51, indicating that almost the entire range of the scale (0–57) was represented.

The difference between the sum-score based on Lyle's decision rules and the sum-score of all

19 items, plotted as a function of the sum-score based on all 19 items, is presented in Figure 1. This graph shows that the Lyle sum-score leads to lower scores in 16 patients (25%), most of whom had a relatively low 19-item sum-score, and to higher scores in 12 patients (19%) with a relatively high 19-item sum-score.

The Mokken scale analysis resulted in a single scale comprising all 19 items, with a scalability coefficient H of 0.79. H per item ranged from 0.69 to 0.86. The reliability coefficient rho was 0.98. A detailed check for monotonicity showed a worst Crit value of 31, indicating that the scale met the assumption of monotone homogeneity. The worst Crit value for nonintersection was 140 for the item Ball bearing, 6 mm, 3rd finger and thumb (item 1 of the subtest Pinch in the Appendix), indicating a violation of the assumption of invariant item ordering. The items with the worst Crit value for nonintersection were removed stepwise, until all Crit values in the remaining item-set were well below 80. This procedure led to subsequent removal of the following items (all from the subtest Pinch): (1) Ball bearing, 6 mm, 3rd finger and thumb (item 1); (2) Marble 3rd finger and thumb (item 5); (3) Ball bearing 2nd finger and thumb (item 3); (4) Ball bearing 1st finger and thumb (item 4). The scalability coeffi-

Table 1 Baseline characteristics of the 63 chronic stroke patients included in the present study

Median age (interquartile range)	61	(52–66)
Median years since stroke (interquartile range)	3.0	(1.9–5.0)
Females	27	(42.9%)
Diagnosis of haemorrhage	16	(25.4%)
Left-sided hemiparesis	11	(17.5%)
Sensory disorders present	28	(44.4%)
Hemineglect present[a]	7	(11.1%)
Mean baseline 19-item ARA sum-score (SD)	30.27	(13.16%)

[a]Information is missing in one case.
SD, standard deviation.

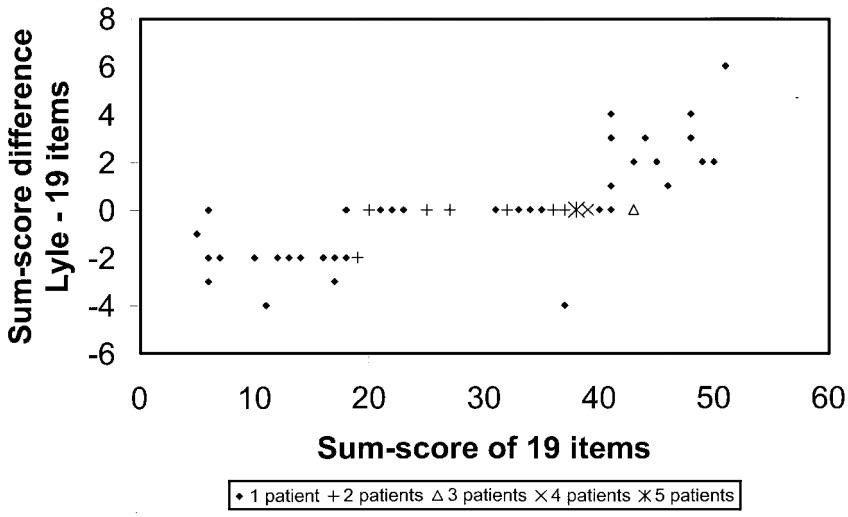

Figure 1 Scatterplot of the difference between the sum-score when using the decision rules proposed by Lyle ('Lyle sum-score') and the sum-score of all 19 items of the ARA test against the 19-item sum-score. Positive differences indicate that the Lyle sum-score is higher, and negative differences indicate that the Lyle sum-score is lower than the 19-item sum-score.

cient H of the resulting 15-item scale was 0.83, H per item ranged from 0.74 to 0.90, and the reliability coefficient rho was 0.97. The mean scores and scalability coefficients H per item of the 15 items are presented in order of ascending difficulty in Table 2.

Comparison of the order of items, as proposed by Lyle (shown in the Appendix), and the empirically derived order of the 15 items in Table 2 makes it clear that the order of items within subtests is very similar, but the items of the different subtests are intermingled. The worst Crit value was 47 for monotonicity and 50 for nonintersection. Therefore, this 15-item scale can be considered to constitute a unidimensional hierarchical scale.

For the assessment of the responsiveness the data from the experimental group ($n = 31$) were used.[15] The means and standard deviations of the two baseline scores and the post-treatment score for each of the three versions of the ARA test, as well as the responsiveness ratios, are presented in Table 3.

As can be seen in this table, the responsiveness ratio is highest (most favourable) for the 19-item and 15-item versions, and lowest for the original

Table 2 Mean scores and scalability coefficient H per item of the unidimensional hierarchical scale of 15 items resulting from the Mokken scale analysis

Item[a]	Mean score[b]	Item H
1) Hand to mouth	2.33	0.76
2) Block 2.5 cm	1.97	0.88
3) Tube 2.25 cm	1.97	0.83
4) Place hand on top of head	1.84	0.83
5) Block 5 cm	1.84	0.88
6) Tube 1 cm	1.83	0.85
7) Stone	1.73	0.86
8) Block 7.5 cm	1.73	0.86
9) Ball 7.5 cm	1.60	0.90
10) Place hand behind head	1.54	0.80
11) Marble 1st finger and thumb	1.49	0.76
12) Pour water from glass to glass	1.49	0.82
13) Washer over bolt	1.38	0.74
14) Marble 2nd finger and thumb	1.35	0.81
15) Block 10 cm	1.35	0.84

[a]Items are arranged in ascending order of difficulty.
[b]Possible scores: 0 = no movement possible; 1 = task partially performed; 2 = task performed, but abnormally; 3 = task performed normally.

Table 3 Means and standard deviations (SD) of the two baseline scores and the post-treatment score of the experimental group ($n = 31$) for each of the three versions of the ARA test, as well as the responsiveness ratios

	Lyle's version		19 items		15 items	
	Mean	SD	Mean	SD	Mean	SD
Baseline 1 (B1)	34.2	13.5	33.7	12.2	28.1	9.6
Baseline 2 (B2)	33.8	11.8	33.4	10.6	27.6	8.1
Post-treatment (P)	39.4	13.9	39.2	13.1	32.6	10.3
Baseline 2 – 1	–0.4	4.7	–0.4	3.4	–0.5	3.0
Post – Baseline 2	5.6	5.3	5.8	5.4	5.0	4.4
Responsiveness ratio[a]	1.2		1.7		1.7	

[a]Responsiveness ratio = $\frac{\text{mean difference P–B2}}{\text{standard deviation of B2–B1.}}$

version using Lyle's decision rules. However, the differences are small.

Discussion

The choice of valid, reliable and responsive outcome measures for clinical trials is often difficult. Instead of developing new outcome measures, existing measurement instruments should be tested empirically, and improved if necessary. In an earlier publication about an intra- and inter-rater reliability study of the ARA test, we recommended the use of time-limits for performance, based on a sample of healthy persons, which followed from the results.[6] These recommended time-limits have been used in the present application of the test.

The wide range of ARA sum-scores represented in the present sample ensures the suitability of this sample for scale analysis. Since the ARA test has distinct ceiling and floor effects in patients with a nearly normal arm function and in patients with a severely impaired arm function, respectively, this type of analysis can only be done in a selected group of patients. Inclusion of patients with lower (i.e. <5) or higher (>50) 19-item sum-scores would not have altered our conclusions. The experimental data obtained from this sample of 63 chronic stroke patients do not support the division of the 19 items of the ARA test into the four subtests proposed by Lyle, since it was found to be a unidimensional scale. Although the division of the 19 tasks into the four subscales grasp, grip, pinch and gross movements is plausible from a practical point of view, the decision rules for skipping should be based on experimental data. It is obvious that the under- and overestimation of respectively low and high sum-scores based on Lyle's decision rules, compared with the sum-scores of the 19 items, will result in a less favourable (i.e. deviating from the Gaussian) distribution of scores in any sample. This supports the use of all items of the ARA test instead of using the decision rules for each separate subtest.

Clinical messages

- The Action Research Arm test is a unidimensional performance test of upper extremity motor function consisting of 19 items.
- There is no rationale for using subtests and decision rules as presented in the original test.
- A subset of 15 items constitutes a hierarchical scale.
- The 15 items in the empirically derived hierarchical order presented here can be used for adaptive testing, thereby substantially reducing testing time.

However, by removing four items, a unidimensional hierarchical scale could be constructed with even higher values of the scalability coefficient H. This 15-item version of the ARA test can be used for adaptive testing, if prior knowledge is available about the patient population (e.g. from a pilot study). Adaptive testing means that only a part of the ARA test needs to be performed, either in the more difficult or the less difficult range of the test, according to the abilities of the patient, thereby substantially reducing the time needed for testing. Especially for patients with a severely impaired arm function, failure to perform the required tasks can be very frustrating. Therefore, adaptive testing not only reduces testing time but it may also prevent patients from becoming frustrated, and even decrease the risk of drop-outs in clinical trials. The tasks can be tried in order of ascending or descending difficulty, starting with Hand to mouth or with Block 10 cm, respectively (Table 2). We propose to move to the next task until four subsequent tasks have yielded the same score (either 0, when the order is ascending, or 3, when the order is descending). In our data-set we could only check this for the ascending order. Two out of nine patients with three subsequent scores of 0 on tasks of increasing difficulty, had a score of 1 on the next task (and after that only scores of 0 for the more difficult tasks).

The four items which were subsequently removed, as described in Results, seem to represent the most difficult tasks in the original test. The fact that the item response functions of these items violated the nonintersection assumption implies that the order of difficulty of these four

items compared with the other items was aberrant. This may be related to the characteristics of the patient sample in this study. As can be seen in Table 1, 44% of the patients had sensory disorders, and it is conceivable that this caused difficulties, in particular with regard to the items involving picking up the small ball bearing with two fingers only. Due to a lack of information about the presence or absence of sensory disorders in the patient sample studied by Lyle, the comparability of the present sample to that in Lyle's study remains uncertain.[1] The generalizability of the present findings remains to be confirmed by replication in different (e.g. less chronic) patient samples.

The proposed changes result in a shorter test, consisting of 15 items, which can be used for adaptive testing, and appears to be more responsive than Lyle's original test. The possible sum-scores range from 0 to 45. There are no reasons to suspect that this empirically derived item-ordering reduces the validity or reliability of the ARA test.

References

1 Lyle RC. A performance test for assessment of upper limb function in physical rehabilitation treatment and research. *Int J Rehabil Res* 1981; **4**: 483–92.

2 Carroll D. A quantitative test of upper extremity function. *J Chronic Dis* 1965; **18**: 479–91.

3 De Weerdt W, Harrison MA. Measuring recovery of arm-hand function in stroke patients: a comparison of the Brunnstrom-Fugl-Meyer test and the Action Research Arm test. *Physiother Can* 1985; **37**: 65–70.

4 Wagenaar RC, Meijer OG, Van Wieringen PC *et al*. The functional recovery of stroke: a comparison between neuro-developmental treatment and the Brunnstrom method. *Scand J Rehabil Med* 1990; **22**: 1–8.

5 Hsieh CL, Hsueh IP, Chiang FM, Lin PH. Inter-rater reliability and validity of the Action Research Arm test in stroke patients. *Age Ageing* 1998; **27**: 107–14.

6 Van der Lee JH, De Groot V, Beckerman H, Wagenaar RC, Lankhorst GJ, Bouter LM. The intra- and inter-rater reliability of the Action Research Arm test: a practical test of upper extremity function in patients with stroke. *Arch Phys Med Rehabil* 2001; **82**: 14–19.

7 Crow JL, Lincoln NB, Nouri FM, De Weerdt W. The effectiveness of EMG biofeedback in the treatment of arm function after stroke. *International Disabil Studies* 1989; **11**: 155–60.

8 Dekker JH, Wagenaar RC, Lankhorst GJ, De Jong BA. The painful hemiplegic shoulder: effects of intra-articular triamcinolone acetonide. *Am J Phys Med Rehabil* 1997; **76**: 43–48.

9 Feys HM, De Weerdt WJ, Selz BE *et al*. Effect of a therapeutic intervention for the hemiplegic upper limb in the acute phase after stroke. A single blind, randomized, controlled multicenter trial. *Stroke* 1998; **29**: 785–92.

10 Broeks JG, Lankhorst GJ, Rumping K, Prevo AJ. The long-term outcome of arm function after stroke: results of a follow-up study. *Disabil Rehabil* 1999; **21**: 357–64.

11 Kwakkel G, Wagenaar RC, Twisk JW, Lankhorst GJ, Koetsier JC. Intensity of leg and arm training after primary middle-cerebral-artery stroke: a randomised trial. *Lancet* 1999; **354**: 191–96.

12 Lincoln NB, Parry RH, Vass CD. Randomized, controlled trial to evaluate increased intensity of physiotherapy treatment of arm function after stroke. *Stroke* 1999; **30**: 573–79.

13 Parry RH, Lincoln NB, Vass CD. Effect of severity of arm impairment on response to additional physiotherapy early after stroke. *Clin Rehabil* 1999; **13**: 187–98.

14 Powell J, Pandyan AD, Granat M, Cameron M, Stott DJ. Electrical stimulation of wrist extensors in poststroke hemiplegia. *Stroke* 1999; **30**: 1384–89.

15 Van der Lee JH, Wagenaar RC, Lankhorst GJ, Vogelaar TW, Devillé WL, Bouter LM. Forced use of the upper extremity in chronic stroke patients: results from a single-blind randomized clinical trial. *Stroke* 1999; **30**: 2369–75.

16 Van der Lee JH, Beckerman H, Lankhorst GJ, Bouter LM. The responsiveness of the Action Research Arm test and the Fugl-Meyer Assessment scale in chronic stroke patients. *J Rehabil Med* 2001; **33**: 110–13.

17 Guttman L. A basis for scaling quantitative data. *Am Soc Rev* 1944; **9**: 139–50.

18 De Souza LH. The development of a scale of the Guttman type for the assessment of mobility disability in multiple sclerosis. *Clin Rehabil* 1999; **13**: 476–81.

19 Deelman BG, Koning-Haanstra M, Liebrand WBG, Van de Burg W. *Manual for the SAN test* [in Dutch]. Lisse: Swets en Zeitlinger, 1987.

20 Folstein MF, Folstein SE, McHugh PR. 'Mini-mental state'. A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* 1975; **12**: 189–98.

21 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **1**: 307–10.

22 Molenaar IW, Sijtsma K. *User's manual MSP5 for Windows. A program for Mokken scale analysis*. Version 5.0. Groningen: iec ProGAMMA, 2000.
23 Mokken RJ. *A theory and procedure of scale analysis*. New York/Berlin: De Gruyter, 1971.
24 Roorda LD, Roebroeck ME, Lankhorst GJ, Van Tilburg T, Bouter LM. Measuring functional limitations in rising and sitting down: development of a questionnaire. *Arch Phys Med Rehabil* 1996; **77**: 663–69.
25 Guyatt GH, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Disabil* 1987; **40**: 171–78.

Appendix – The ARA subtests and decision rules for skipping items as presented by Lyle[1]

All items are rated on a four-point ordinal scale (0 = no movement possible, 3 = task performed normally).

Subtest Grasp	1) Block 10 cm (If score = 3, total for subtest Grasp = 18 and proceed to subtest Grip) 2) Block 2.5 cm (If score = 0, total = 0 and proceed to subtest Grip) 3) Block 5 cm 4) Block 7.5 cm 5) Ball 7.5 cm 6) Stone
Subtest Grip	1) Pour water from glass to glass (If score = 3, total for subtest Grip = 12 and proceed to subtest Pinch) 2) Tube 2.25 cm (If score = 0, total for subtest Grip = 0 and proceed to subtest Pinch) 3) Tube 1 cm 4) Washer over bolt
Subtest Pinch	1) Ball bearing, 6 mm, 3rd finger and thumb (If score = 3, total for subtest Pinch = 18 and proceed to subtest Gross movements) 2) Marble 1st finger and thumb (If score = 0, total for subtest Pinch = 0 and proceed to subtest Gross movements) 3) Ball bearing 2nd finger and thumb 4) Ball bearing 1st finger and thumb 5) Marble 3rd finger and thumb 6) Marble 2nd finger and thumb
Subtest Gross movements	1) Place hand behind head (If score = 3, total for subtest Gross movements = 9 or if score = 0, total for subtest Gross movements = 0) 2) Place hand on top of head 3) Hand to mouth